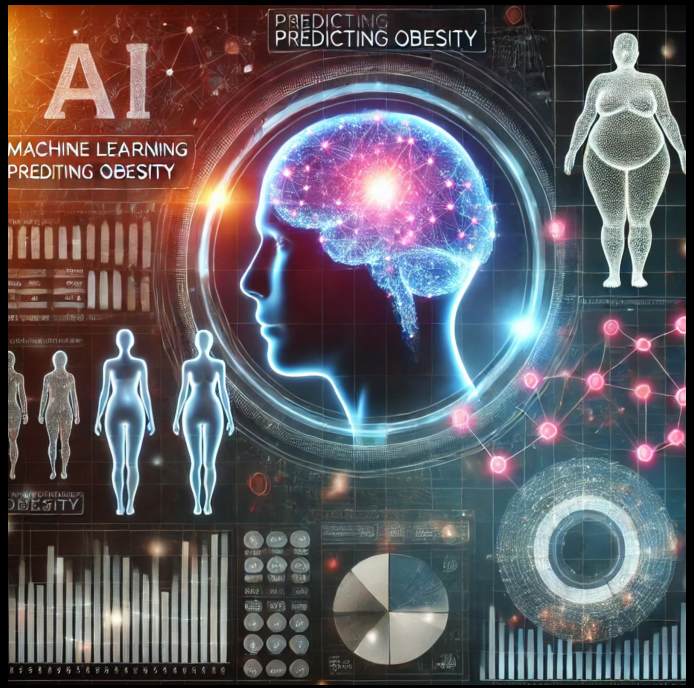


Predicting Obesity with Machine Learning

Elham Rahimi

Introduction

- Obesity is a global health concern with many medical implications
- Early predictions can help prevent obesity
- Machine learning lets us use data to predict obesity risk based on lifestyle and eating habits.

Project Objectives

- To develop a machine learning model to predict the likelihood of a person becoming obese
- Identify factors that contribute to the likelihood of a person becoming obese
- With this model; meal planners, diet centers and gyms can use this model to help their customers.

Usage

Meal Planning Services

Diet Centers

Gym

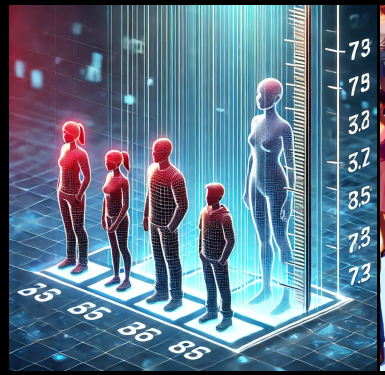
Doctors

Dataset

- Kochar, J. P. (n.d.). *Obesity Risk Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/jpkochar/obesity-risk-dataset

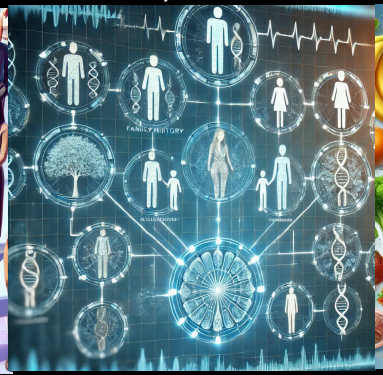
The features in the dataset include:

Anthropometric Features



Demographic Features

Genetic & Family History Features



Dietary Habits & Nutrition Features

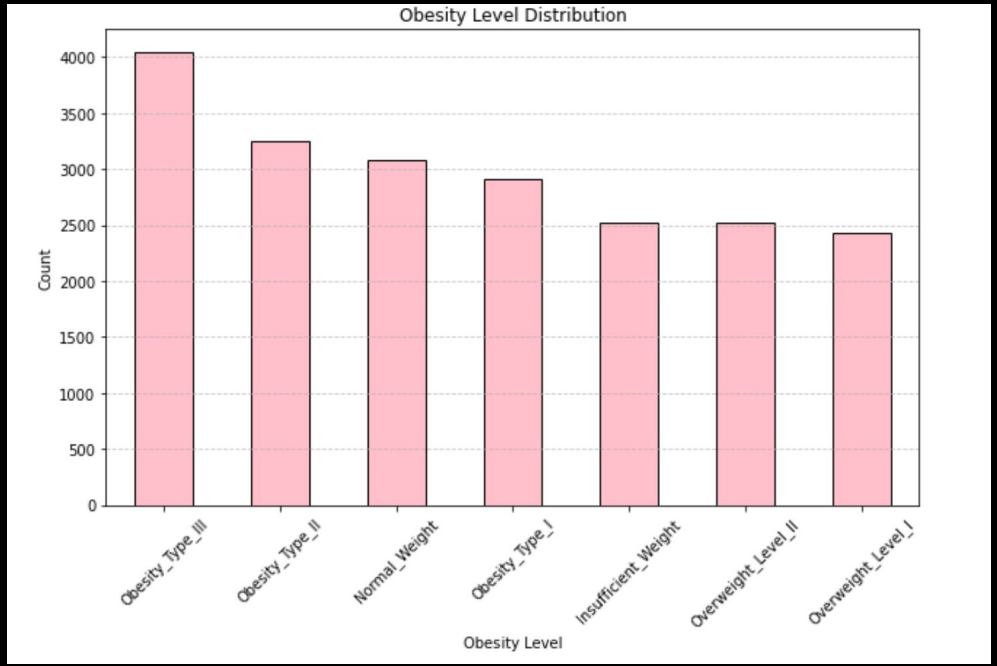
Lifestyle Features

Data Dictionary

Feature	Description
Gender	Gender of the individual (Male/Female)
Age	Age in years
Height	Height in meters
Weight	Weight in kilograms
family_history_with_overweight	Whether there is a family history of being overweight (Yes/No)
FAVC	Frequent consumption of high-calorie food (Yes/No)
FCVC	Frequency of vegetable consumption (1 = rarely, 2 = sometimes, 3 = always)
NCP	Number of main meals per day
CAEC	Consumption of food between meals (No, Sometimes, Frequently, Always)
SMOKE	Whether the individual smokes (Yes/No)
CH2O	Daily water intake in liters
SCC	Whether the person monitors caloric intake (Yes/No)
FAF	Physical activity frequency (hours per week)
TUE	Daily screen time (hours per day)
CALC	Frequency of alcohol consumption (No, Sometimes, Frequently)
MTRANS	Main mode of transportation (e.g. Walking, Bike, Public Transportation)
obesity_level	Target variable: Obesity classification (7 categories)

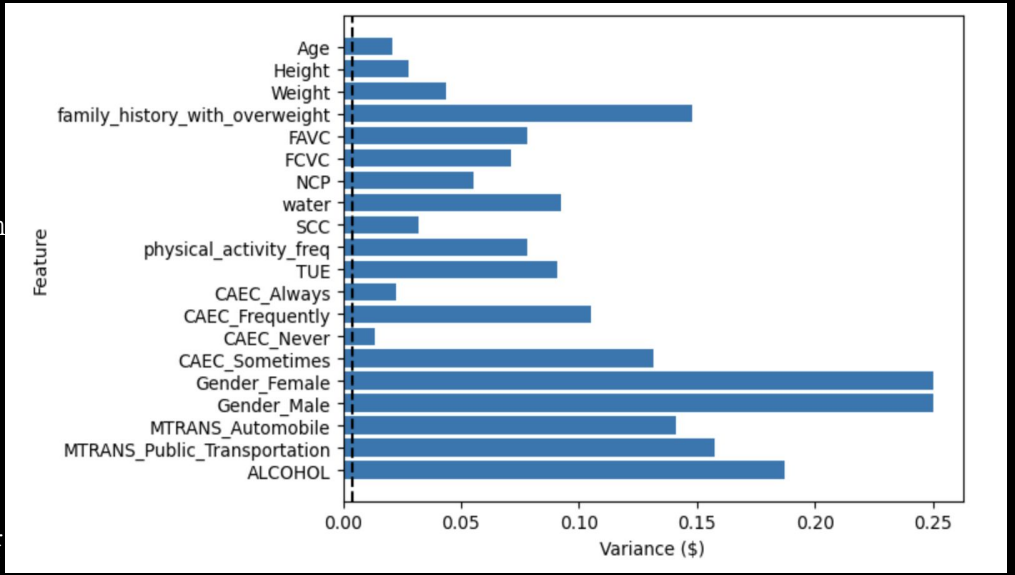
Distribution of Obesity Level by BMI

Obesity Level	Estimated BMI
Insufficient Weight	〈 18.5
Normal Weight	18.5 – 24.9
Overweight Level I	25 – 29.9
Overweight Level II	30 – 34.9
Obesity Type I	35 – 39.9
Obesity Type II	40 – 44.9
Obesity Type III	〉45



EDA and Feature Engineering

- One-hot encoded all categorical features.
- Dropped low-frequency categories (e.g., Smoking: 1.18%, Walking: 2.25%, Two-Wheeler: 0.34%) to reduce noise.
- Transformed skewed features like Age to improve distribution
- Removed low-variance features after scaling (threshold = 0.004).
- Dropped multicollinear features **only for Logistic Regression** (not needed for tree-based models).
- Applied PCA to reduce dimensionality while retaining 95% of variance.
- Also explored grouped PCA for visualization and comparison.



Modeling Approach

Trained and compared three models:

- Logistic Regression
  - Strong Base Model
  - Works with linear features
  - Sensitive to multicollinearity
  - Fast to train
  - Easy to interpret
- Decision Tree
  - No need for scaling
  - Can easily overfit
  - Handles non-linear features
  - Easy to visualize and explain
  - Sensitive to small changes
- Random Forest
  - Handles overfitting
  - Captures complex patterns
  - Slow to train
  - Handles high dimensional data
  - Less outliers to noise and outliers

Model Evaluation

	precision	recall	f1-score	support
Insufficient_Weight	0.89	0.94	0.91	524
Normal_Weight	0.86	0.82	0.84	626
Obesity_Type_I	0.83	0.84	0.84	543
Obesity_Type_II	0.96	0.97	0.97	657
Obesity_Type_III	1.00	1.00	1.00	804
Overweight_Level_I	0.73	0.74	0.73	484
Overweight_Level_II	0.73	0.69	0.71	514
accuracy			0.87	4152
macro avg	0.86	0.86	0.86	4152
weighted avg	0.87	0.87	0.87	4152

Performs well overall with consistent precision and recall.

Strong on distinct classes like Obesity_Type_III (F1-score: 1.00).

Struggles slightly on borderline categories like Overweight_Level_I & II.

High performance on distinct classes (Obesity_Type_II & III).

Better than Logistic Regression in handling overlapping classes.

Can overfit without tuning but gives fast and clear decisions.

	precision	recall	f1-score	support
0	0.93	0.90	0.92	524
1	0.83	0.84	0.83	626
2	0.74	0.70	0.72	484
3	0.74	0.79	0.76	514
4	0.86	0.84	0.85	543
5	0.97	0.96	0.96	657
6	0.99	1.00	0.99	804
accuracy			0.88	4152
macro avg	0.86	0.86	0.86	4152
weighted avg	0.88	0.88	0.88	4152

WINNER!

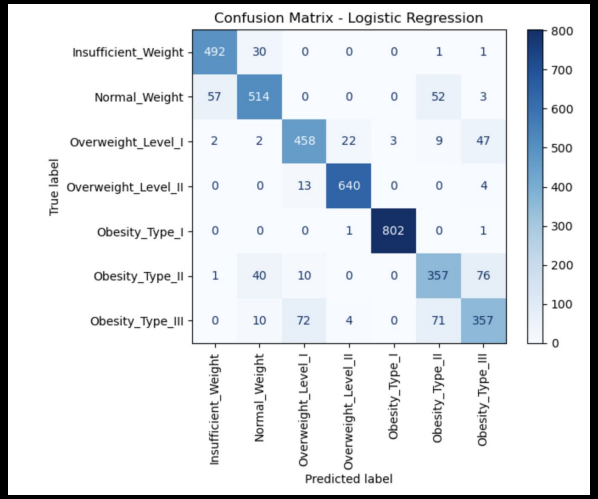
	precision	recall	f1-score	support
Insufficient_Weight	0.94	0.93	0.94	524
Normal_Weight	0.87	0.88	0.87	626
Obesity_Type_I	0.88	0.88	0.88	543
Obesity_Type_II	0.97	0.97	0.97	657
Obesity_Type_III	1.00	1.00	1.00	804
Overweight_Level_I	0.77	0.77	0.77	484
Overweight_Level_II	0.80	0.81	0.81	514
accuracy			0.90	4152
macro avg	0.89	0.89	0.89	4152
weighted avg	0.90	0.90	0.90	4152

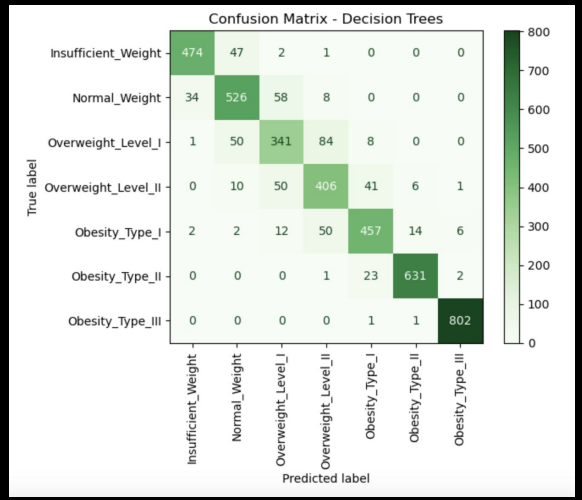
Best performance overall across all metrics.

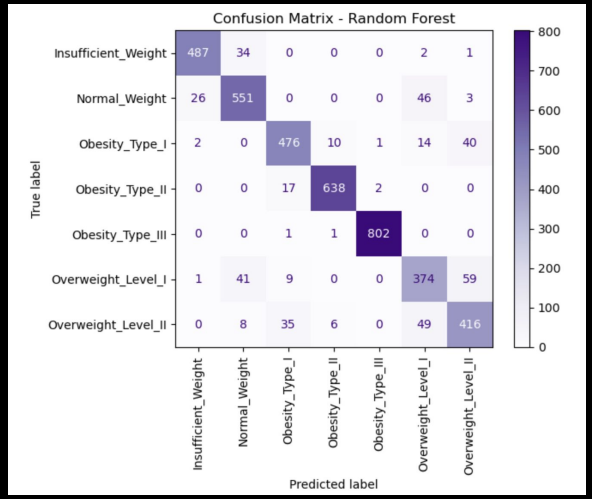
Very strong F1-scores across all classes, including difficult ones.

Handles complexity and class overlap better due to ensemble structure.

Confusion Matrix







Winner!

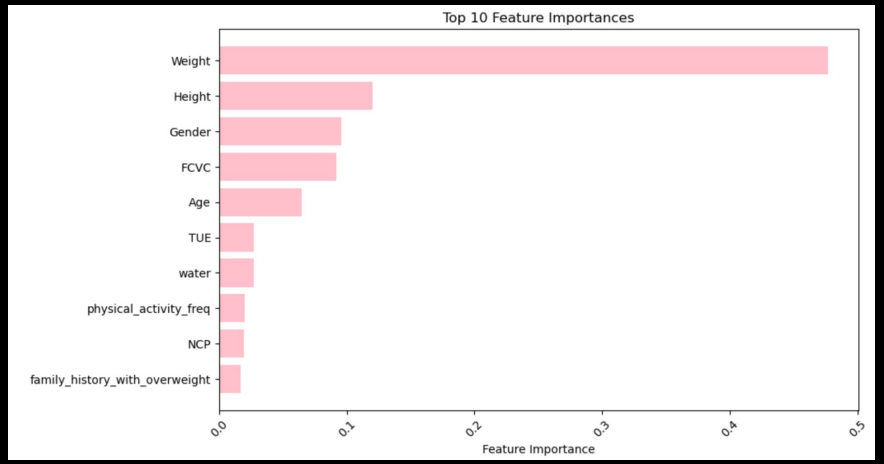
Highest accuracy (90%)

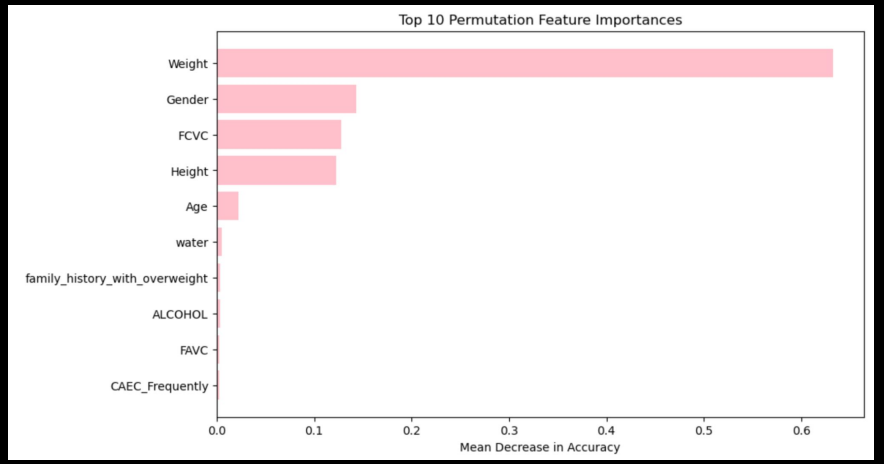
Strongest consistency across all classes

Least confusion in middle-ground categories

Best generalization without overfitting

Model Interpretation





- Weight: Strongest predictor by far, model accuracy drops sharply when shuffled, central to obesity classification
- Gender: The model relies heavily on gender, probably because male/female obesity thresholds differ. May also reflect differences in fat distribution, lifestyle habits, or health behavior.
- FCVC (Frequency of Vegetable Consumption) Surprisingly impactful. Suggests that healthier eating habits (like regularly eating vegetables) are significant markers in distinguishing weight classes.
- Height: Likely interacts with weight to help the model infer BMI-like relationships. While not directly indicative of obesity alone, it's critical when used with weight.
- Age: Minor but non-negligible contribution. Older individuals may show patterns of weight gain or reduced activity, but it’s not a major standalone driver in this model.

Next steps | Ways to improve

- Remove obvious features and focus on what features that people can actually control in that moment. Retrain model to look at features related to diet, exercise, alcohol, and water.
- That will be called Habit–Based Obesity Risk Estimator
- Create personalized recommender system like “You're at risk of Obesity Type I. Based on your data, try increasing veggie intake (FCVC) and drinking more water daily. Also, reducing snacks between meals (CAEC) could help.”
- Add more features like cultural background.

*Better to **prevent** than treat. This model helps people do that*